LEXICAL FUNCTIONS AND MACHINE TRANSLATION

Dirk Heylen, Kerry G. Maxwell and Marc Verhagen

OTS, Trans 10, 3512 JK Utrecht, Netherlands
CLMT Group, Essex University, Colchester, Essex CO4 3SQ, England
email: heylen@let.ruu.nl, kerry@essex.ac.uk, verhm@essex.ac.uk

cmp-lg/9410009

This paper discusses the lexicographical concept of *lexical functions* ([Mel'čuk and Žolkovsky, 1984]) and their potential exploitation in the development of a machine translation lexicon designed to handle collocations. We show how lexical functions can be thought to reflect cross-linguistic meaning concepts for collocational structures and their translational equivalents, and therefore suggest themselves as some kind of language-independent semantic primitives from which translation strategies can be developed.[1]

1 Description of the Problem

Collocations present specific problems in translation, both in human and automatic contexts. If we take the construction *heavy smoker* in English and attempt to translate it into French and German, we find that a literal translation of *heavy* yields the wrong result, since the concept expressed by the adjective (something like `'excessive'`) is translated by *grand* (large) in French and *stark* (strong) in German. We observe then that in some sense the adjectives *stark*, *grand* and *heavy* are *equivalent* in the collocational context, but that this is of course not typically the case in other contexts, cf *grande boite, starke Schachtel* and *heavy box*, where the adjectives could hardly be viewed as equivalent. It seems then that adjectives which are not literal translations of one another may share meaning properties specifically in the collocational context.

How then can we specify this special equivalence in the machine translation dictionary? The answer seems to lie in addressing the concept which underlies the union of adjective and noun in these three cases, i.e., intensification, and hence establish a single meaning representation for the adjectives which can be viewed as an interlingual pivot for translation.

Collocations
have been studied by computational linguists in different contexts. For instance, there is a substantial body of papers on the *extraction* of "frequently co-occurring words" from corpora using statistical methods (e.g., ([Choueka *et al.*, 1983]), ([Church and Hanks, 1989]), ([Smadja, 1993]) to list only a few). These authors focus on techniques for providing material that can be used in other processing tasks such as word sense disambiguation, information retrieval, natural language generation and so on. Also, the *use* of collocations in different applications has been discussed by various authors (([McRoy, 1992]), ([Pustejovsky *et al.*, 1992]),
([Smadja and McKeown, 1990]) etc.). However, collocations are not only considered useful, but also a *problem* both in certain applications (e.g. generation, ([Nirenburg *et al.*, 1988]), machine translation, ([Heid and Raab, 1989])) and from a more theoretical point of view (e.g. ([Abeillé and Schabes, 1989]), ([Krenn and Erbach, To appear])).

We have been concerned with investigating the *lexical functions* (LFs) of Mel'čuk ([Mel'čuk and Žolkovsky, 1984]) as a candidate interlingual device for the translation of adjectival and verbal collocates. Our work is related to research by ([Heid and Raab, 1989]). In some respects it is an extension of some of their suggestions. Our work differs from theirs in scope and also in the exploration of various other directions.

2 Representation

The use we make of lexical functions as interlingual representations, does not respect their original Mel'čukian interpretation. Furthermore, we have transferred them from their context in the Meaning-Text Theory to a different theoretical setting. We have embedded the concept in an HPSG-like grammar theory.[2] In this section we review this operation. First we consider the features of Mel'čuk's treatment that we have wanted to preserve. Next we show how they have been imported into the HPSG framework.

[1]The research reported in this paper was undertaken as the project "Collocations and the Lexicalisation of Semantic Operations" (ET-10/75). Financial contributions were by the Commission of the European Community, Association Suissetra (Geneva) and Oxford University Press.

[2]Head Driven Phrase Structure grammar, see ([Pollard and Sag, 1987]), ([Pollard and Sag, to appear]). For another treatment of collocations in HPSG, see ([Krenn and Erbach, To appear]).

2.1 Collocations and LFs

In Mel'čuk's *Explanatory Combinatory Dictionary* (ECD, see ([Mel'čuk *et al.*, 1984])), expressions such as *une ferme intention, une résistance acharnée, un argument de poids, un bruit infernal* and *donner une leçon, faire un pas, commetre un crime* are described in the lexical combinatorics zone. These "expressions plus ou moins figées" will be called 'collocations'. They are considered to consist of two parts — the *base* and the *collocate*. In the examples above, the nouns are the bases and the adjectives and the verbs are the collocates. The idea that all adjective collocates and all the verb collocates share an important meaning component — roughly paraphrasable as *intense* and *do* respectively — and the fact that the adjectives and verbs are not interchangeable but are restricted with this meaning to the accompanying nouns, is coded in the dictionary using lexical functions (in this case **Magn** and **Oper**).

Each article in the ECD describes what is called a 'lexeme': a word in some specific reading. In the *lexical combinatorics zone*, we find a list of the lexical functions that are relevant to this particular lexeme. Each lexical function is followed by one or more lexemes (the result or value of the function applied to the head word). The idea is that each combination of the argument with one of the values of the function forms a collocation in our terminology. The argument corresponds to the base and each value is a collocate. The following features of this representation are important to us.

- Lexical functions are used to represent an important syntactico-semantic relation between the base and the collocate.
- The restricted combinatorial potential of the collocate lexeme is accounted for by listing it at each base with which it can occur.

The second of these characteristics points out that the collocational restriction is seen as a purely lexical, idiosyncratic one: all collocations are explicitly listed.

One other aspect of collocations which we have to deal with is the relation between the collocate lexeme and its freely occurring counterpart. Collocate lexemes often differ in some respects from their literal variants while sharing other properties. Mel'čuk deals with this by including in the ECD an entry for the free variant and putting the collocate-specific information in the entry for the base (with the result of the lexical functions). The full entry of the collocate is the result of taking the entry for the free variant and overwriting it with the information provided at the base.

2.2 Collocations in HPSG

The three aspects of Mel'čuk's analysis we wanted to encode in HPSG were the following.

- Coding the base-collocate relation in the lexicon.
- Choosing the level at which lexical functions will be situated.
- Relating the collocate information to the free variant entry.

We have provided straightforward solutions to these problems. For the first problem we have taken over the ECD architecture rather directly, by creating a dedicated 'collocates' field in the entry for bases which contains all the relevant collocates. As far as the second problem is concerned, the obvious place to put lexical functions is in the semantic representation provided by HPSG. There are various reasons for this. One is that LFs are used in the deep syntax level in Mel'čuk's model, a level oriented towards meaning. Another reason is that this level seems most appropriate to be used in transfer/translation and because we want to use lexical functions in transfer, this is where they should be. In contrast to the ECD, the meaning of the collocate is represented by the lexical function only.

The following is an example of the entry for *criticism* with the encoding of *strong* as a collocate.[3] We use SEM_IND as an abbreviation for the feature path SEM.CONT.IND.

$$\left[\begin{array}{ll} \text{PHON} & \text{criticism} \\ \text{SEM\_IND} & \left[\begin{array}{ll} \text{VAR} & \boxed{1} \\ \text{REST} & \{\text{criticism}(\boxed{1})\} \end{array}\right] \\ \text{COLLS} & \{\left[\begin{array}{ll} \text{\$strong} & \\ \text{SEM\_IND} & \left[\begin{array}{ll} \text{VAR} & \boxed{1} \\ \text{REST} & \{\textbf{Magn}(\boxed{1})\} \end{array}\right] \end{array}\right]\} \end{array}\right]$$

Just as in the ECD the base contains a specific zone in which the collocates are listed. In our case, the feature 'COLLS' has a set of lexical entries as its value.

Each collocate subentry bears the value of the lexical function in its semantics field. In this representation the lexical function is chosen as the real semantic value of the collocate. One should read the feature structure as specifying that the semantics of *strong* (as a collocate) is the predicate **Magn**(1).

The collocate subentry only provides partial information. In fact, it provides only the information that is specific to the occurrence of *strong* in its combination with *criticism*. In this case only the semantics is given. We further assume that the lexicon also contains a 'super-entry' which provides all the information that is shared by all the different occurrences of *strong*. This entry is where the variable $strong points to. Of course, other architectures that try to avoid redundant specification of information are equally possible. For instance if one assumes a mechanism of default unification,

[3] Notice that here we use a simple version of HPSG based on ([Pollard and Sag, 1987]) whereas the actual implementation was based on ([Pollard and Sag, to appear]).

one can have $strong refer to the full entry describing 'strong' in say its ordinary use, and have the values that are particular to the collocational *strong* overwrite the values provided in the ordinary entry, as in Mel'čuk's proposal.

Collocations, Rules and Principles So far, we have not specified in what way one gets from the lexical entries for the base and the collocate to the representation of the collocational expression.

In HPSG, the descriptions of complex expressions are constrained by principles. We will assume that collocations are subject to the same constraints. The ordinary rules of combination (combining adjectives and nouns, for instance) thus account for most of the properties of the collocational combination. However, we are still left with the typical 'collocational restriction' which needs to be accounted for.

We have therefore added a principle which says that constructions that are analysed as collocations (indicated by the type COLLOCATION) are either head-adjunct structure or head-complement structures with specific restrictions holding between the head and the adjunct or the head and the complement respectively. Let's consider the former case[4], illustrated by the *heavy smoker* example. The adjunct daughter will contain the adjective collocate. In such collocational constructions the collocate adjuncts have to be 'licensed' by the noun or the head daughter. This is implemented by requiring that the collocates field (COLLS) of the head daughter contains a reference to a lexical entry that is compatible with the adjunct daughter. In the literal reading of an expression such as *heavy smoker*, the phrase will not be analysed as a COLLOCATION and the principle does not apply.

$$\begin{bmatrix}\ \end{bmatrix}_{\text{COLLOCATION}} \Rightarrow \begin{bmatrix}\text{HEAD\_DTR} & \begin{bmatrix}\text{COLLS} & \{\ldots\boxed{1}\ldots\}\end{bmatrix} \\ \text{ADJ\_DTRS} & <\ldots\boxed{1}\text{COLLOCATE}\cdots>\end{bmatrix}$$

$$\begin{bmatrix}\text{HEAD\_DTR} & \boxed{1}\text{COLLOCATE} \\ \text{COMP\_DTRS} & <\ldots[\text{COLLS}\{\ldots\boxed{1}\ldots\}]\ldots>\end{bmatrix}$$

3 Issues in Translation

The project has tried to investigate the use of lexical functions as an interlingual device, i.e., one which is shared by the semantic representations of collocations in the language pairs[5].

[4]To illustrate the case of head-complement structures one could take some support verb construction (also called light verb construction).

[5]For another application of LFs in a multilingual NLP context see ([Heid and Raab, 1989]). For other treatments of collocations in language generation see ([Nirenburg *et al.*, 1988]) and ([Smadja and McKeown, 1990]).

The typing of a collocation with such a function opens up the way to a treatment of collocations inside a given language module and hence to a substantial reduction in the number of collocations explicitly handled in the multilingual transfer dictionary. The existence of a collocation function is established during analysis. This information is used to generate the correct translation in the target language. To illustrate, the English analysis module might analyse (1) as (2). The transfer module maps (2) onto (3) which is then synthesised by the French module to (4).

(1) heavy smoker → (2) **Magn**(smoker) → (3) **Magn**(fumeur) → (4) grand fumeur

The example points out that the translation strategy is a mixture of transfer and interlingua. The bases are transferred but the representation of the collocate is shared between the source and the target representation. This treatment of collocations rests, among others, on the assumptions that there are only a limited number of lexical functions, that lexical functions can be assigned consistently, that all (or a significant number of) collocations realise a lexical function, that lexical functions are not restricted to particular languages, etc. In the following paragraph we present an outline of the translation process. Next, we discuss some of the problems which follow from our approach and we propose some ways to solve them.

3.1 Lexical Functions as Interlingua

It was assumed that the starting point for transfer is the semantic representation of the phrase. Using a semantic representation as input to transfer implies that we relate semantic values of words and phrases. For our purposes this is very satisfying since we will now be using the semantics of collocates instead of their orthography, in other words: we use lexical functions and abstract away from the particular realisation of a collocate in a particular language.

We now state the relation between the semantic representations of the source language and target language. The semantic relation between the phrase *heavy smoker* and its French counterpart can be made explicit in the following bilingual sign:

$$\begin{bmatrix}\text{EN—SEM\_IND} & \begin{bmatrix}\text{VAR} & \boxed{1} \\ \text{REST} & \{\text{smoker}(\boxed{1}),\mathbf{Magn}(\boxed{1})\}\end{bmatrix} \\ \text{FR—SEM\_IND} & \begin{bmatrix}\text{VAR} & \boxed{1} \\ \text{REST} & \{\text{fumeur}(\boxed{1}),\mathbf{Magn}(\boxed{1})\}\end{bmatrix}\end{bmatrix}$$

Typically, the lexicon will contain a bilingual sign for each possible value of RELN. Thus, for translating *heavy smoker* into *grand fumeur* we will need the obvious entry for *smoker-fumeur* plus the entry below:

$$\left[\begin{array}{l} \text{EN—SEM\_IND} \left[\begin{array}{ll} \text{VAR} & \boxed{1} \\ \text{REST} & \{\textbf{Magn}(\boxed{1})\} \end{array}\right] \\ \\ \text{FR—SEM\_IND} \left[\begin{array}{ll} \text{VAR} & \boxed{1} \\ \text{REST} & \{\textbf{Magn}(\boxed{1})\} \end{array}\right] \end{array}\right]$$

The interlingual status of the lexical function is self-evident. Any occurrence of **Magn** will be left intact during transfer and it will be the generation component that ultimately assigns a monolingual lexical entry to the LF.[6]

3.2 Problems

Lexical Functions abstract away from certain nuances in meaning and from different syntactic realizations. We discuss some of the problems raised by this abstraction in this section.

Overgenerality An important problem stems from the interpretation of LFs implied by their use as an interlingua — namely that *the meaning of the collocate in some ways reduces to the meaning implied by the lexical function.* This interpretation is trouble-free if we assume that LFs always deliver unique values; unfortunately cases to the contrary can be readily observed. An example attested from our corpus was the range of adverbial constructions possible with the verbal head *oppose*: *adamantly, bitterly, consistently, steadfastly, strongly, vehemently, vigorously, deeply, resolutely*, etc. The function **Magn** is an appropriate descriptor in all cases since each adverb functions as a typical intensifier in this context. However each adverb also denotes some other meaning aspect(s). The imprecision of LFs will mean that we have no means of distinguishing between the various intensifiers possible in the context of a given keyword, and hence will not have sufficient information to choose the most appropriate translation where, correspondingly, multiple possibilities exist in the target language. An important question here is how dramatic this loss of translation quality really is.

It is essentially in addressing the issue of overgenerality that Mel'čuk introduces sub- and superscripts to lexical functions, enhancing their precision and making them sensitive to meaning aspects of the lexical items over which they operate. Superscripts are intended to make the meaning of the LF more precise and hence more likely to imply unary mappings between arguments and values, subscripts are used to reference a particular semantic component of a keyword. The introduction of such devices into the account of LFs demonstrates both the need for precision and the fact that it does seem necessary to address semantic aspects of lexemes standing in cooccurrence relations. In fact it has been asserted by some (e.g., ([Anick and Pustejovsky, 1990]), ([Heid and Raab, 1989])) that collocational systems are systematically predictable from the lexical semantics of nouns. In an attempt to explore this notion further, we have investigated the approach to nominal semantics known as *Qualia* structure ([Pustejovsky, 1991]) and considered how this may complement the LF notion to improve its descriptive power[7]. Among the promising avenues that occur to us are, firstly, the postulation of LF subscripts based on the four Qualia roles (assuming that these are the lexically most relevant aspects of noun semantics) and, secondly, the application of LFs to semantic (Qualia) structures rather than monolithic lexemes; eg: the LF **Bon** is used in delivering evaluative qualifiers which are standard expressions of praise or approval. One could imagine application of the function over the Constitutive and Agentive roles of the noun *lecture*, to deliver:

Bon(`Const:` lecture) = informative
Bon(`Agent:` lecture) = clear

In both cases the idea is that the precision of the lexical function is essentially enhanced by appealing to the semantic facets of its argument.

Syntactic Divergences Another issue that has to be raised concerns the translation of collocations into non-collocational constructions. If we are to maintain a consistent interlingual approach to the translation of these cases, we must extend our LF-based approach accordingly. We consider one case briefly.

Cross-linguistic analysis reveals many cases where nominal-based collocational constructs are realised as compounds in Germanic languages, e.g., *bunch of keys* ⇒ *sleutelbos.* A possible account of such phenomena may be developed from the concept of *merged* LFs ([Mel'čuk and Žolkovsky, 1970]). Merged LFs are intended to be used in cases where a value lexeme exists which appears to effectively reduce ("merge") an LF meaning and its specified argument to a single lexicalised form, rather than projecting a syntagmatic unit. We could argue that in cases of compound formation, exactly the same process is to be accounted for, since the compound embodies both the concept mediated by the LF and its argument lexeme. We could therefore allow compounds to be delivered as values of merged LF's, eg: `//`**Mult**(sleutel)= sleutelbos.

These observations are useful in the MT context if we assume that we can effect a mapping between merged and unmerged LFs and therefore capture the correspondence between distinct structural realisations of the same concept. One way to emulate such a mapping might be through the use of

[6] For more details we refer the reader to ([Heylen, 1993]). There we also discuss our implementation in Alep, the C.E.C.'s unification-based grammar writing environment.

[7] For a comparison between aspects of Qualia structures and lexical functions see ([Heylen, to appear]).

Mel'čuk's *lexical paraphrasing rules*. For instance, one could conceive of a lexical paraphrasing rule as follows[8]: $W + \mathbf{Mult}(W) \Longleftrightarrow //\mathbf{Mult}(W)$.

If we assume that in our monolingual English lexicon, we assign the collocate *bunch* as the **Mult** value of keyword *key*, and that accordingly in the Dutch lexical entry for *sleutel* we instantiate *sleutelbos* as the value of the merged LF **//Mult**, then we can use the paraphrasing rule to effect a mapping between the two LF's and hence arrive at an interlingual approach to the translation of the example, despite structural mismatches, i.e.,

$$\begin{array}{c} \text{key} + \text{bunch}[\mathbf{Mult}(key)] \\ \Longleftrightarrow \\ \text{sleutelbos}[//\mathbf{Mult}(sleutel)] \end{array}$$

Further examples exist where productive morphological processes (e.g., affixation[9]) lead to the lexicalisation in one language of concepts that exist as syntagmatic constructs in another. Again, we suggest the use of merged LFs and corresponding mappings via lexical paraphrasing rules as a possible translation strategy in these cases.

4 Summary and Conclusions

In this paper we have discussed how the lexicographical concept of *lexical functions*, introduced by Mel'čuk to describe collocations, can be used as an interlingual device in the machine translation of such structures. We have shown how the essentials of the ECD analysis can be embedded in the lexicon and grammar of a unification based theory of language.

Our use of lexical functions as an interlingua assumes that the relevant aspects of the meaning of the collocate are fully captured by the LF. The LF therefore determines the accuracy of translations, which may be impoverished due to the generalised nature of basic LFs. We have suggested some ways in which LFs can be enriched with lexical semantic information to improve translation quality.

The interlingua level reflects what is semantically common to expressions which form translational equivalents. It abstracts away from specific syntactic realisations. Given that collocations may translate as non-collocations, we also have to provide a way to represent these expressions using lexical functions. We have provided an illustration on how to proceed in one such case.

[8]This is our own initiative – it seems to be the case as we examine the literature that neither LFs such as **Magn**, **Bon** etc (i.e., those representing standard qualifiers/attributes) nor indeed *merged* LFs feature in lexical paraphrasing rules. We would argue that cross-linguistic analysis suggests that they should enter this domain; compound formation and other types of lexicalisation appear to be regular patterns of translation across many collocational constructs, as we illustrate here.

[9]One could think of an example such as *mis-interpret*.

Acknowledgements We would like to thank the following partners and colleagues: Susan Armstrong-Warwick, Laura Bloksma, Nicoletta Calzolari, R. Lee Humphreys, Simon Murison-Bowie and André Schenk.

References

[Abeillé and Schabes, 1989] A. Abeillé and Y. Schabes. Parsing idioms in lexicalized tags. In *EACL/89*, Manchester, 1989.

[Anick and Pustejovsky, 1990] P. Anick and J. Pustejovsky. An application of lexical semantics to knowledge acquisition from corpora. In *Coling/90*, Helsinki, 1990.

[Choueka *et al.*, 1983] K. Choueka, S.T. Klein, and E. Neuwitz. Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *ALLC Journal*, pages 34–38, 1983.

[Church and Hanks, 1989] K. W. Church and P. Hanks. Word association norms, mutual information and lexicography. In *ACL/89*, Vancouver, 1989.

[Heid and Raab, 1989] U. Heid and S. Raab. Collocations in multilingual generation. In *EACL/89*, pages 130–136, Manchester, 1989.

[Heylen, 1993] Dirk Heylen. Collocations and the lexicalisation of semantic operations. Technical report, OTS, 1993.

[Heylen, to appear] D. Heylen. Lexical functions and knowledge representation. In P. Saint-Dizier and E. Viegas, editors, *Computational Lexical Semantics*. CUP, to appear.

[Krenn and Erbach, To appear] B. Krenn and G. Erbach. Idioms and support verb constructions. In J. Nerbonne, K. Netter, and C. Pollard, editors, *German Grammar in HPSG*. CSLI Lecture Notes, To appear.

[McRoy, 1992] S. W. McRoy. Using multiple knowledge sources for word sense discrimination. *Computational Linguistics*, 18(1):1–30, 1992.

[Mel'čuk and Žolkovsky, 1970] I.A. Mel'čuk and A.K. Žolkovsky. Sur la synthèse sémantique. *T.A. Informations*, 2:1–85, 1970.

[Mel'čuk and Žolkovsky, 1984] I.A. Mel'čuk and A.K. Žolkovsky. *Explanatory Combinatorial Dictionary of Modern Russian*. Wiener Slawistischer Almanach Sonderband 14, Vienna, 1984.

[Mel'čuk *et al.*, 1984] I. A. Mel'čuk, N. Arbatchewsky-Jumarie, L. Elnitsky, L. Iordanskaja,

and A. Lessard. *Dictionnaire explicatif et combinatoire du français contemporain*. Les Presses de l'Université de Montréal, Montreal, 1984.

[Nirenburg *et al.*, 1988] S. Nirenburg, R. McCardell, E. Nyberg, S. Huffman, E. Kenschaft, and I. Nirenburg. Lexical realization in natural language generation. In *Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Pittsburgh, 1988.

[Pollard and Sag, 1987] C. Pollard and I. Sag. *Information Based Syntax and Semantics*. CSLI, Stanford, 1987.

[Pollard and Sag, to appear] C. Pollard and I. Sag. Head driven phrase structure grammar. to appear.

[Pustejovsky *et al.*, 1992] J. Pustejovsky, S. Bergler, and P. Anick. Lexical semantic techniques for corpus analysis. *Computational Linguistics*, 19(2):331–358, 1992.

[Pustejovsky, 1991] J. Pustejovsky. The generative lexicon. *Computational Linguistics*, 17(4), 1991.

[Smadja and McKeown, 1990] Frank Smadja and Kathleen R. McKeown. Automatically extracting and representing collocations for language generation. In *28th Annual Meeting of the Association for Computational Linguistics*, Pittsburgh, PA, 1990.

[Smadja, 1993] F. Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177, 1993.